

# Multisets

Luciano da Fontoura Costa

*luciano@ifsc.usp.br*

*São Carlos Institute of Physics – DFCM/USP*

15th Oct 2021

## Abstract

Multisets are sets that allow repetition of elements, therefore account for their frequency of observation. As such, multisets pave the way to a number of interesting possibilities of both theoretical and applied nature. In the present work, after revising the main aspects of traditional sets, we introduce some of the main concepts and characteristics of multisets, which is followed by their generalization to take into account vectors, matrices, functions, scalar fields, and densities. A new approach is also proposed in which the negative multiplicities are allowed, implying the multiset universe to become finite and well-defined, corresponding to the multiset with all support element associated with null multiplicities. It then becomes possible to define the complement operation in multisets in a robust manner, which allows properties involving complement – including the De Morgan theorem – to be recovered in multisets. The possibility to define set operation between mfunctions that are analogous to the traditional inner product is also proposed, paving the way to obtaining respective mfunction transformations analogous to the Fourier transform. This result also allowed the proposal of performing hybrid signal processing operations on mset mfunctions, including filtering and template matching. Relationships between the cosine similarity index and the Jaccard index are also identified, including the presentation of an intersection-based variation of the cosine index. The potential of multisets in pattern recognition and deep learning is also briefly characterized and illustrated, more specifically regarding the frequent issue of comparing clusters or densities.

‘In the bag, seashells gathered long ago resound.’

---

*LdaFC*

## 1 Introduction

Multisets can be informally understood as sets capable of incorporation of repeated entries of the same element (e.g. [1, 2, 3, 4, 5, 6]). In a sense, they are at least as compatible with human experience than sets. For instance, when one goes to the greengrocer to buy oranges, it does matter if our bag contains 3 or 4 elements.

In the present work, we aim at providing an introduction to this interesting area, while also briefly covering the Jaccard index adapted to multiset, and applications to pattern recognition. Some new results are also described, including the possibility to allow negative multiplicities, allowing the multiset universe to be composed of null multiplicities for all support elements. This allows the complement operation to become stable, recovering analogous properties to all counterparts in traditional sets, including the De Morgan theorem.

We start by reviewing the main concepts and properties of traditional sets, and then present the concept of multisets, as well as some of their simpler properties, also including several examples. The challenges implied by the definition of a universe set for multisets is briefly characterized and discussed. It is also argued that the operations of sum and subtraction between multisets correspond to one of the main distinction between multiset and set theories.

The possibility to generalize multisets to several other mathematical structures including vectors, matrices, functions, scalar and vector fields, as well as probability densities are approached next, including several examples. In particular, the extension of multisets as representations of functions and scalar fields paves the way for obtaining hybrid expressions involving combinations of the set operations of union and intersection with algebraic expressions involving sum, subtraction, product and division of sets.

The interesting possibility to obtain an analogue of the traditional inner product functional, as well as respectively based transformations, is addressed next, with the introduction of the concepts of mproducts and common

products, which leads to the understanding of the Jaccard index for functions as a normalized version of the common product. The relationship between the cosine similarity index and the Jaccard index is also addressed from the perspective of multisets and multifunctions.

We then presents how the definition of the common product between two msets or mfunctions paves the way to obtaining hybrid signal operations including filtering and template matching.

The Jaccard index, as well as its extension to multisets and multiple arguments, is then briefly presented as an interesting manner to compare any of the mathematical structures mentioned above after they have been transformed into respective multisets.

The application of multisets and the Jaccard index to quantify the relationship between two or more clusters is then described with respect to an example related to the iris dataset. The measurement of the separation between clusters corresponds to an important issue in both pattern recognition (e.g. [7, 8]), deep learning (e.g. [9, 10, 11]), and modeling (e.g. [12, 13])

For simplicity's sake, the term multisets are henceforth abbreviated as *msetmset*.

## 2 Traditional Sets

A *set* is an unordered collection of items, or *elements*, which are not allowed to repeat. A set  $A$  with elements  $a$ ,  $b$ , and  $c$  is typically represented as:

$$A = \{a, b, c\}$$

The two essential properties of sets therefore are that the elements may appear in any order, which distinguish sets from vectors, and that the elements cannot be repeated.

The number of elements in a set is called its *cardinality* or *size*, being represented as  $|A|$ .

A *subset*  $B$  of a given set  $A$  consists of a set so that any of its elements are contained in  $A$ . If  $A$  contains  $N$  elements, there will be  $|A| = 2^N$  possible subsets that can be derived from it. The set containing all possible subsets of  $A$  is called its *power set*  $P^A$ .

An important point about sets that is sometimes overlooked regards the fact that they always refer to a respective *universe set*  $\Omega$ . More specifically, once this set is established, any possible set needs to be a subset of  $\Omega$ . Observe that  $\Omega$  can have any type of element, though the situation where the elements are homogenous is of particular interest.

In case some sets are given but the universe set is not provided, it is still possible to estimate the respective universe set (within hypotheses and subject to incomplete-

ness in case new sets appear) as corresponding to the union of the supplied sets.

The universe set is of fundamental importance because the operation of *complement* of a set is defined with respect to the universe set. More specifically, the complement of a set  $A$  consists of all elements of  $\Omega$  that are not part of  $A$ . The complement of a set is henceforth represented as  $A^C$ , being implicit that the operation refers to a given  $\Omega$ .

Sets can be finite or infinite, as well as discrete or continuous. A *finite set* is any set  $A$  so that  $|A| < \infty$ . A *discrete set* is characterized by having all its elements corresponding to isolated points  $p$ , in the sense that each of these points possesses a neighborhood which when united with the universe set yields only  $p$ . Any continuous set is infinite, but discrete sets can be finite or infinite.

An interesting point regards the relationship between an element, let's say ' $a$ ' and the set  $\{a\}$ . These two mathematical structures are not identical because it is possible to include an element into  $\{a\}$ , but not into ' $a$ '.

The *empty set*, represented as  $\phi = \{\}$  is a subset of any possible set.

Given two sets  $A$  and  $B$ , their *union* consists of a third set  $C$  containing all elements from  $A$  and  $B$ . The *intersect* of these two sets corresponds to a set  $C$  containing all elements that are in both  $A$  and  $B$ . A subset  $B$  of  $A$  can therefore be understood as to be so that  $A \cap B = B$ . Any set is a subset of itself.

The *difference* between two sets  $A$  and  $B$ , indicated as  $A - B$ , corresponds to the set  $C$  containing all elements that are in  $A$  but are not in  $B$ .

Given three sets  $A$ ,  $B$ , and  $C$  derived from a given  $\Omega$ , the following properties are satisfied:

$$A \cup A^C = \Omega \quad (1)$$

$$A \cap A^C = \phi \quad (2)$$

$$A \cup \phi = A \quad (3)$$

$$A \cap \phi = \phi \quad (4)$$

$$A \cup A = A \quad (5)$$

$$A \cap A = A \quad (6)$$

$$A \cup B = B \cup A \quad (7)$$

$$A \cap B = B \cap A \quad (8)$$

$$A \cup (B \cap C) = (A \cup B) \cap C \quad (9)$$

$$A \cap (B \cup C) = (A \cap B) \cup C \quad (10)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad (11)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \quad (12)$$

$$(A \cup B)^C = A^C \cap B^C \text{ (De Morgan)} \quad (13)$$

$$(A \cap B)^C = A^C \cup B^C \text{ (De Morgan)} \quad (14)$$

### 3 Multisets

Basically, *msets* are sets allowing the *repetition* of elements, which is understood as their *multiplicity* or *frequency*. As with sets, the order of the elements is immaterial. Examples of mset include:

$$\begin{aligned} A &= \{a, a, b, b, b, d\}; \\ B &= \{1, 2, 1, 2, 1, 2, 1\} = \{1, 1, 1, 1, 2, 2, 2\}; \\ C &= \{1, a, 2, b, b, 3, c, c, c, 1, d, 2, a, a\} = \\ &= \{1, 1, 2, 2, 3, a, a, a, b, b, c, c, c\}; \\ D &= \{a, a, b, d\}. \end{aligned}$$

Observe the different symbol adopted henceforth in this work in order to emphasize the distinction between a traditional set ( $\{\}$ ), and a mset ( $\{\!\!\{\}$ ).

A more compact representation of a mset  $A$  can be obtained by using 2-tuple or pairs  $[a, m(a)]$ , where ‘ $a$ ’ is an element and  $m(a)$  its multiplicity, i.e. the number of times it appear in  $A$ . In the case of the above examples, we have:

$$\begin{aligned} A &= \{a, a, b, b, b, d\} = \{[a, 2]; [b, 2]; [d, 1]\}; \\ B &= \{1, 1, 1, 1, 2, 2, 2\} = \{[1, 4]; [2, 3]\}; \\ C &= \{1, 1, 2, 2, 3, a, a, a, b, b, c, c, c\} = \\ &= \{[1, 2]; [2, 2]; [3, 1]; [a, 3]; [b, 2]; [c, 3]\}; \\ D &= \{a, a, b, d\} = \{[a, 2]; [b, 1]; [d, 1]\}. \end{aligned} \quad (15)$$

Though this type of representation of msets actually corresponds to a set, because it is impossible to have two identical entries, we shall maintain the ‘ $\{\!\!\{\}$ ’ notation in order to emphasize that a mset is being meant.

When referring to the multiplicity of an element, it is important to specify to which mset this is being referred. This can be done by writing  $m_A(a)$ , meaning the multiplicity of the element  $a$  in the mset  $A$ .

The property analogous to inclusion in sets can be stated as follows. A mset  $A$  is included in another mset  $B$  whenever  $m_A(a) \leq m_B(a)$ . For instance, in the case of the examples above, we have  $m_A(a) = 2$  and  $m_C(a) = 2$ .

As with sets, it is particularly important to specify the universe of a mset. This can be done in an analogous manner as with sets. As an example, let’s obtain a possible universe set for the two msets  $A$  and  $B$  above:

$$\Omega = \{a, b, d, 1, 2\} \quad (16)$$

It should be kept in mind that this universe is not analogous to the counterpart in sets, as it does not actually account for the possible multiplicity of the involved elements, which is unbound through the sum operation.

It is now possible to rewrite those two sets in a more

complete, though redundant manner, as follows:

$$\begin{aligned} P &= \{[a, 2]; [b, 2]; [d, 1]; [1, 0]; [2, 0]\}; \\ Q &= \{[a, 0]; [b, 0]; [c, 0]; [1, 4]; [2, 3]\}; \end{aligned}$$

The *support* of a given mset  $A$  is defined as:

$$S_A = \{x | x \in \Omega, m(x) > 0\} \quad (17)$$

For instance, the supports of the msets in Equation ??

$$\begin{aligned} S_A &= a, b \\ S_B &= 1, 2 \\ S_C &= 1, 2, a, b, c \\ S_D &= a, b, d \end{aligned} \quad (18)$$

As such, this set can be understood as containing all distinct elements in  $A$ . Observe that the support set provides a useful index for identifying the possible elements in the respective msets.

### 4 Multiset Operations

A set  $A$  is said to be *included* into another set whenever:

$$m_A(x) \leq m_B(x), \forall x \in A \quad (19)$$

For simplicity’s sake, we will indicate this operation using the same symbol as for sets, i.e.  $A \subset B$ , as the type of operation can be inferred from  $A$  and  $B$  being sets or msets.

In the case of the mset examples above, we can write that  $D \subset A$ .

The *union*  $C$  of two msets  $A$  and  $B$  can be defined as:

$$\begin{aligned} C &= A \cup B = \{[x, m_C(x)], x \in A \text{ or } x \in B\}, \\ &\text{with } m_C(x) = \max\{m_A(x), m_B(x)\} \end{aligned} \quad (20)$$

Examples considering the msets in the beginning of Section 3 include:

$$\begin{aligned} A \cup B &= \{a, a, b, b, b, d, 1, 1, 1, 1, 2, 2, 2\} = \\ &= \{[1, 2]; [b, 2]; [d, 1]; [1, 4]; [2, 3]\} \\ A \cup D &= \{a, a, b, d\} = \{[a, 4]; [b, 4]; [d, 2]\} \end{aligned}$$

It is interesting to observe that the resulting multiplicity of each element does *not* correspond to the *sum* of the respective multiplicities, but to the *maximum* between them. This is a particularly important point that deserves further contemplation, so we will be back to it after presenting the concept of *sum* of two msets.

Let  $A$  and  $B$  be msets. The *sum* of these two sets, henceforth represented as  $C = A + B$ , is defined as:

$$\begin{aligned} C &= A + B = \{[x, m_C(x)], x \in A \text{ or } x \in B\}, \\ &\text{with } m_C(x) = m_A(x) + m_B(x) \end{aligned} \quad (21)$$

Figure 1 illustrates the two different ways in which the common elements of two msets  $A$  and  $B$  are collected into their respective union and sum msets.

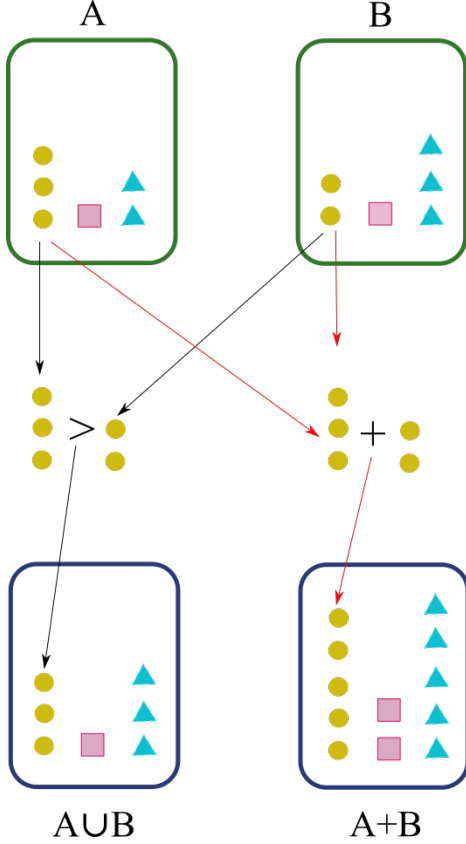


Figure 1: The union (a) and sum (b) of two msets  $A$  and  $B$  typically yield quite different resulting msets. In the case of the union operation, each of the elements of the same type are compared, with the elements with the maximum multiplicity being incorporated into  $C$ . The sum of the two msets incorporates all the  $m_A(x_i) + m_B(x_i)$  elements into  $C$ .

Examples respective to the msets in the beginning of Section 3 include:

$$A + B = \{a, a, b, b, b, d, 1, 1, 1, 1, 2, 2\} = \{[1, 2]; [b, 2]; [d, 1]; 1, 4; [2, 3]\}$$

$$A + D = \{a, a, a, a, b, b, b, b, d, d\} = \{[a, 4]; [b, 4]; [d, 2]\}$$

Thus, we have that the mset operations of union and sum are related in the sense that both collect the elements from the two msets, but the way in which this is done is quite different, with the multiplicities of the mset obtained by union becoming necessarily smaller or equal than that of the mset obtained by sum, i.e.  $m_{A \cup B}(x_i) \leq m_{A+B}(x_i)$ .

It is interesting to consider these two operations in the context of possible respective applications. The sum of the two msets ensures conservation of the total number

of elements (such as in conservative or flow-related problems), being therefore more indicated for related situations. The union of two msets can be conceptually understood as a kind of mid point between the sum of msets and the conventional union of traditional sets. Though the union of msets will typically yield larger msets than the union, it will not guarantee conservation of the total number of elements. A typical situation in which the union of msets can be applied is when the incorporation of the elements from the two msets is performed in terms of a choice, with the larger set being taken.

The *intersection* between two msets  $A$  and  $B$  can be defined as:

$$C = A \cap B = \{[x, m_C(x)], x \in A \text{ or } x \in B\},$$

$$\text{with } m_C(x) = \min\{m_A(x), m_B(x)\} \quad (22)$$

Examples drawn from the msets in the beginning of Section 3 include:

$$A \cap B = \{[a, 4]; [b, 4]; [d, 2]\}$$

$$A \cap D = \{a, a, b\} = \{[a, 4]; [b, 4]; [d, 2]\} \quad (23)$$

The *difference* or *subtraction* between two msets is expressed as:

$$C = A - B = \{[x, m_C(x)], x \in A \text{ or } x \in B\},$$

$$\text{with } m_C(x) = \max\{m_A(x) - m_B(x), 0\} \quad (24)$$

It is interesting to observe that the restriction of not having negative values can be overlooked without great impact on the other properties and operations as addressed in the present work. Actually, in Section 7, we will show that allowing negative multiplicities paves the way to a robust definition of the universe mset as well as closed subtraction operations.

Figure 2 illustrates the intersection and difference between two msets  $A$  and  $B$ .

Respective examples include:

$$A - D = \{b, b\} = \{[b, 2]\}$$

$$D - A = \{\} \quad (25)$$

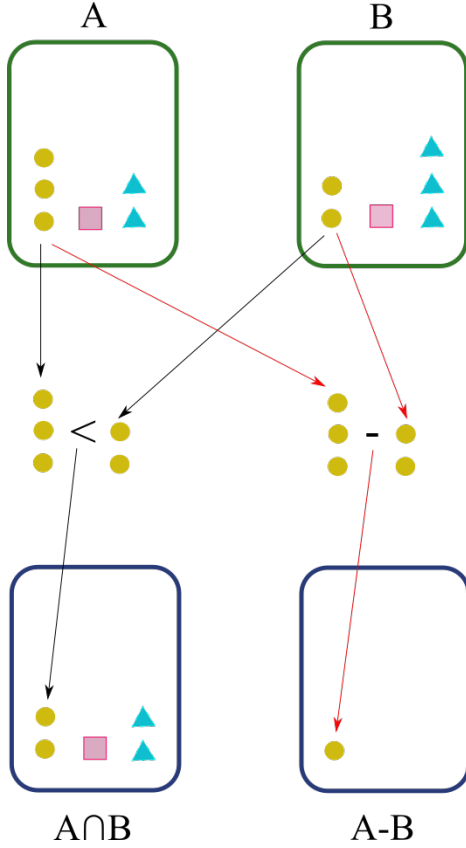


Figure 2: The intersection (a) and difference (b) of two multisets  $A$  and  $B$  typically yield quite different resulting multisets. In the case of the intersection operation, each of the elements of the same type are compared, with the elements with the minimum respective multiplicity being incorporated into  $C$ . The difference between  $A$  and  $B$  depends on  $m_A(x_i) - m_B(x_i)$ . As the result is negative in the case of the present example, no elements are incorporated into  $A - B$ .

## 5 Multisets Properties

It can be shown that multisets satisfy the following properties:

$$A \cup \phi = A \quad (26)$$

$$A \cap \phi = \phi \quad (27)$$

$$A \cup A = A \quad (28)$$

$$A \cap A = A \quad (29)$$

$$A \cup B = B \cup A \quad (30)$$

$$A \cap B = B \cap A \quad (31)$$

$$A \cup (B \cup C) = (A \cup B) \cup C \quad (32)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad (33)$$

$$A \cap (B \cap C) = (A \cap B) \cap (A \cap C) \quad (34)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \quad (35)$$

Informally speaking, the above properties can be understood by identifying each repeated element in an multiset with unique respective tags, in which case they would

behave with respect to the operation of union and intersection exactly in the same way as common sets.

So, we have that multisets follow all the properties in Equations 1 to Equations 14, except those involving complementation. Indeed, the definition of the complement of an multiset has been a challenging issue (e.g. [6]), which has to do with the fact that there is no bound to the size of possible multisets generated by making additions between any non-empty multiset. For instance, we can write:

$$\begin{aligned} A &= \{a\} \\ A &= A + A = \{a, a\} \\ A &= A + A = \{a, a, a, a\} \\ &\dots \end{aligned} \quad (36)$$

Interestingly, as will be presented in Section 7, the universe multiset actually correspond to an multiset with all multiplicities equal to zero. When applied to functions represented as multisets, this means that the null function is the respective universe.

For this reason, the useful De Morgan properties, as well as other related results, do not hold for multisets. In addition, there are relative few properties involving the sum and subtraction of multisets. However, when negative values are allowed for the subtraction between multisets, the universe multiset consists of taking all the multiplicities as zero, and the complement operation becomes the change of sign of the multiplicities (see Equation 40).

In a sense, it is these two operations that differentiates multisets from sets because, as commented above, multisets behave like sets respectively to their union and intersection when the elements are tagged.

Though we gain two new operations when working with multisets, these operations imply some specific challenges to be addressed. In the present work, we will simply assume that the complement of multisets are not available.

## 6 Multisets, Vectors, Matrices

In this section we will discuss the relationship between multisets and vectors. First, we recall that the elements in a vector are expected to follow a well-determined order as indicated by their indices. For instance, in the case of the vector in  $\mathbb{R}^5$ :

$$\vec{v} = [3, 2.5, \pi, 0, -1]$$

we have five indices  $i = 1, 2, \dots, 5$ , so that we can specify the respective element values as  $v[1] = 3$ ,  $v[2] = 2.5$ ,  $v[3] = \pi$ ,  $v[4] = 0$ , and  $v[5] = -1$ .

By understanding the values of the components of a vector as generalized multiplicities, we can immediately derive the following multiset from the above vector:

$$V = \{[1, 3]; [2, 2.5]; [3, \pi]; [4, 0]; [5, -1]\}$$

Therefore, we have that an mset can be derived from any vector, but that a vector can be obtained from an mset only if their elements are ordered in some manner, e.g. by taking their respective values instead of understanding them as labels. This situation becomes more evident when one considers non-numeric elements. As such, msets can be used to study the elements of vectors without taking into account their relative position along the vector.

It is also interesting to contemplate the relationship between the above discussion and the traditional sets containing multiplicities. More specifically, we can write the set containing all multiplicities in the vector  $\vec{v}$  above as:

$$\tilde{V} = \{3, 2.5, \pi, 0, -1\} = \{0, 3, 2.5, -1, \pi\} = \text{etc.}$$

Though  $V$  and  $\tilde{V}$  are very similar, they are not identical because in  $V$  the *correspondence* of the elements and the respective multiplicity is maintained. This difference becomes critically important in case we want to apply the operations of addition and subtraction between msets, which would be otherwise impossible in case of sets because we would not know which element of one set should be added to which element in the other set.

By representing vectors as msets, we not only preserve the operations of subtraction and difference, but also incorporates the possibilities of defining intersections and unions between any two vectors.

Another interesting possibility is to incorporate new operators for multiplication and division into the mset framework, which can be done straightforwardly, while avoiding divisions by zero.

Interestingly, it is also possible to obtain mset representations from matrices or even other more sophisticated mathematical structures as tensors. In the case of matrices, this can be done by mapping the indices  $i = 1, 2, \dots, N_i$  and  $j = 1, 2, \dots, N_j$  into a single index  $k$ , e.g.:

$$k \longleftrightarrow N_i(j - 1) + i - 1 \quad (37)$$

so that the matrix becomes a vector, which can then be transformed into the respective mset as described above. Interestingly, observe that though the obtained msets would not directly provide the respective indices of the elements, they can be nevertheless recovered from the unified index.

## 7 Functions and Scalar Fields

The possibility to represent vectors as msets opens the way to a number of interesting possibilities. One of them is to represent discrete and continuous *functions* and *scalar fields* (vector fields can be approached as vec-

tors of scalar fields). We develop these possibilities in the following.

We start with a discrete function as illustrated in Figure 3.

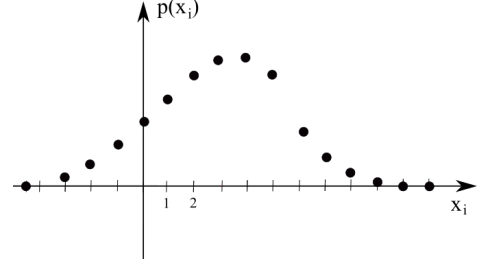


Figure 3: A generic discrete function  $f(x_i)$ .

It is often interesting to represent such discrete functions in terms of sums of Dirac delta functions.

Provided that we limit the extension of this function along the horizontal axis (related to the function support), e.g.  $i = 1, 2, \dots, N$ , it can be immediately represented in terms of the vector:

$$\vec{f} = [f(x_i)] \quad (38)$$

which can then be transformed into the respective mset as described in the previous section.

A similar approach can be applied to transform discrete scalar fields defined on more than one variables into respective msets, involving the index mapping described above.

Now, we can approach the case of continuous functions and scalar fields simply by taking the separation between the points along the horizontal axis to the limit of 0. As a consequence, the functions and scalar fields will become associated to msets with infinite elements, but these can still be operated by the ‘max()’, ‘min()’, ‘+’ and ‘-’ operations required by the mset operations, as well as any other function or transformation applicable to functions. For simplicity’s sake, we shall call the multisets derived from functions as *mfunctions*.

Observe that, though all information about a function is preserved into the respective mfunction, allowing its reconstruction, the adjacency between any two points along the domain is lost in the mfunction because the order of the elements in a mset does not matter.

Let’s illustrate the above concepts and possibilities in terms of the two following functions:

$$\begin{aligned} g(x) &= e^{-10x^2} \\ h(x) &= 2e^{-10|x-0.1|} \end{aligned} \quad (39)$$

Figure 5 illustrates the two above functions as well as



their union, intersection, sum, and subtraction after having been converted into respective msets.

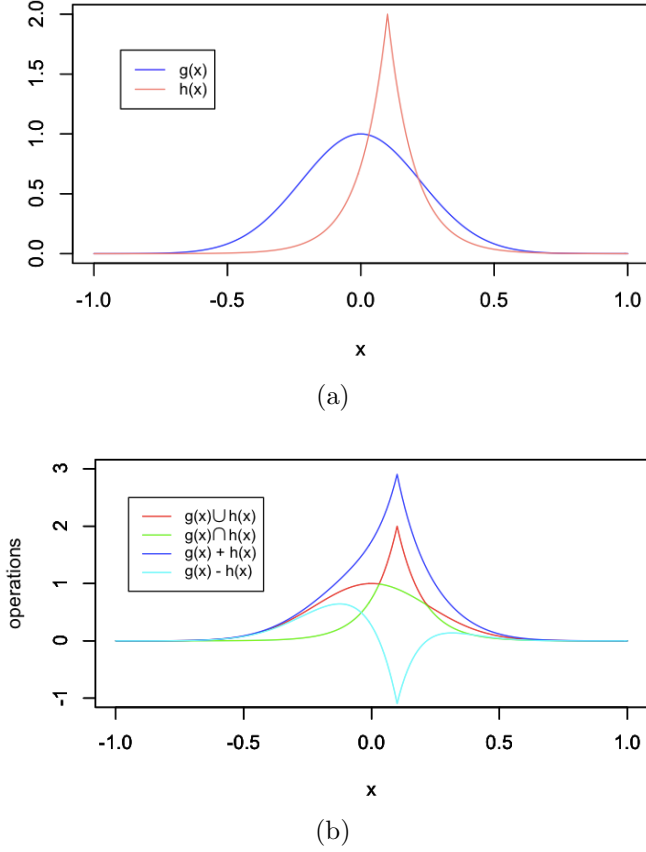


Figure 4: Two continuous functions  $g(x)$  and  $h(x)$  of a single variable (a), and their respective mset operations (b) of union, intersection, sum, and subtraction.

This examples illustrate several interesting points. First, we have the transformation of functions into respective msets immediately allow them to be operated by union and intersection. In addition, we observe an example that the sum of two functions is larger or equal than their union, as well as the possibility of the subtraction operation yielding negative values.

Figure 5 shows the two additional operations of product and quotient between the two functions  $g(x)$  and  $h(x)$  above, avoiding the divisions by zero.

It is interesting to observe that the potential of the operations of union and intersection in producing sharp derivatives and discontinuities, which contribute an interesting manner of representing an ample range of function types as combinations of these operations, not to mention the operations of sum, subtraction, product and quotient.

Consequently, it becomes an interesting possibility to develop transformations of functions, analogous to the

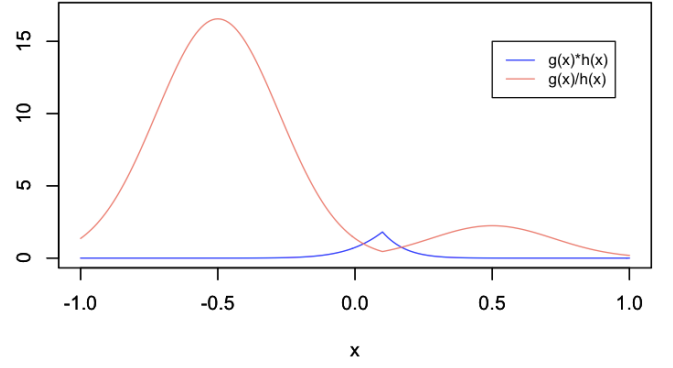


Figure 5: Two functions  $g(x)$  and  $h(x)$  of a single variable (a), and their respective mset operations (b) of union, intersection, sum, and subtraction.

Fourier transform, considering not only series of basis functions, but also intersections and/or unions, and/or other possible operations between msets. One possible interesting benefit would be to become able to express functions with discontinuous derivatives as combinations of functions that are completely smooth. Also, it should be observed that the operations of sum and subtraction are bilinear, while the minimum and maximum are not.

Indeed, the above developments also allow new functions to be obtained as combinations of logic operations as the union and intersection and numeric operations as sum, subtraction, product, and division. For instance, it becomes possible to write things such as:

$$\begin{aligned} r(x) &= (g(x) \cap h(x)) + g(x) \\ s(x) &= (g(x) + h(x)) \cup (g(x) - h(x)) \\ t(x) &= [g(x) \cap h(x)] - [g(x) \cup h(x)] \end{aligned}$$

These three functions are illustrated in Figure 6 assuming the function in Equation 39.

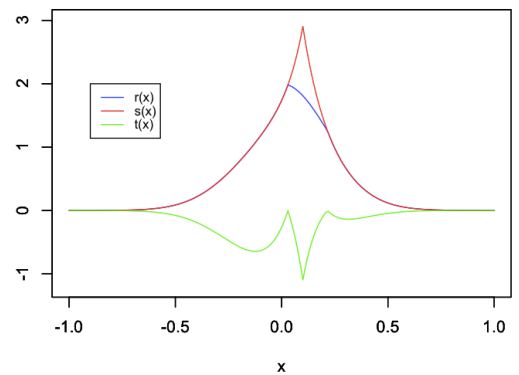


Figure 6: The functions  $r(x)$ ,  $s(x)$  and  $t(x)$  obtained through mset operations.

Seen from this perspective, the complement of a multiset becomes the operation of sign change, in the sense that  $r(x)^C$  would become  $-r(x)$ . Indeed, it can be verified that the De Morgan properties hold in this case, i.e:

$$\begin{aligned} -[g(x) \cap h(x)] &= -g(x) \cup -h(x) \\ -[g(x) \cup h(x)] &= -g(x) \cap -h(x) \end{aligned}$$

The potential of these hybrid constructions is as large as our imagination, because it allows the incorporation of much of the concepts, structures, and properties of both arithmetic, set theory, and also logic (which is directly analogous to set theory).

It should also be observed that the above results could only be achieved from allowing negative values for the multiplicity in the difference operation between msets. The sign change therefore recovers many properties analogous to those involving complements of sets, and more. As a consequence, the universe mset becomes as follows:

$$\Omega = \{[x_1, 0]; [x_2, 0]; \dots; [x_N, 0]\} \quad (40)$$

where  $x_i$  are the elements of the support of the msets.

Other properties analogous to those of traditional sets are also consequently recovered, including  $A \cup A^C$ .

## 8 The Multiset Jaccard Indices

The Jaccard index represents an effective and conceptually appealing manner to quantify the similarity between any two sets  $A$  and  $B$  (e.g. [14, 15, 16]), having therefore being extensively applied in a vast range of problems in several scientific and technological fields.

In its most basic form, the Jaccard index between an two sets  $A$  and  $B$  can be expressed as:

$$\mathcal{J}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (41)$$

It is possible to adapt the Jaccard index to msets by making:

$$\mathcal{J}_M(A, B) = \frac{\sum_{i=1}^N \min(m(a_i), m(b_i))}{\sum_{i=1}^N \max(m(a_i), m(b_i))} \quad (42)$$

where  $a_i$  and  $b_i$  are the elements of the sets  $A$  and  $B$ , respectively, and  $N$  is the cardinality of the universe of those sets. We also have that  $0 \leq \mathcal{J}_M(A, B) \leq 1$ .

As an example, let's consider  $A = \{a, b, b, c, c, c\}$  and  $B = \{a, b, c, c, d\}$ . Then, we have:

$$\mathcal{J}(A, B) = \frac{1 + 1 + 2 + 0}{2 + 3 + 5 + 1} = \frac{4}{11} \quad (43)$$

It is possible to adapt the Jaccard index to mfunctions by making:

$$\mathcal{J}(A, B) = \frac{\int_{\Phi} \min(m_A(\vec{x}), m_B(\vec{x}))}{\int_{\Phi} \max(m_A(\vec{x}), m_B(\vec{x}))} \quad (44)$$

where  $\Phi$  is the common support of the two functions or scalar fields.

As such, the Jaccard index can be understood as a *functional*, or *mfunctional* of the two functions of scalar fields.

The Jaccard index has been enhanced and extended to functions, scalar fields, joint variations and more than 2 sets [16]. In particular, the latter type of Jaccard index for 3 sets  $A$ ,  $B$  and  $C$  can be written as:

$$\mathcal{J}(A, B, C) = \frac{A \cap B \cap C}{A \cup B \cup C} \quad (45)$$

## 9 Inner Product and Transforms

The concept of multifunctions as proposed above has some interesting implications. In particular, it effectively endows mfunctions and scalar fields with the analogous of the set operations of union and intersection, which immediately motivates several important questions such as what are the properties involving both set and algebraic properties of mfunctions. Another possibility of particular relevance because of its immediate relationship with the concepts of similarity and transformation regards if it is possible to define a set-relate operation between two mfunctions which would be analogous to the *inner product*. In this section we propose one such operation and then use it to define transforms of mfunctions in a manner similar to Fourier transform, but without the property of orthogonality extending over all distinct sinusoidals.

We start by recalling the inner product between two real mfunctions  $f(x)$  and  $g(x)$ :

$$\langle f(x), g(x) \rangle = \int_{-\infty}^{\infty} f(x)g(x)dx \quad (46)$$

Any two functions  $f(x)$  and  $g(x)$  are said to be *orthogonal* if and only if their inner product is zero. The prototypical example of orthogonal functions consists in sinusoidals  $s(x, \omega, \phi) = \sin(\omega x + \phi)$ , with  $\omega = 2\pi f$ . It is known that:

$$\langle f(x), g(x) \rangle = \int_{-\infty}^{\infty} s(x, \omega, \phi) s(x, \omega, \tilde{\phi}) dx = 0 \quad (47)$$

provided  $\phi \neq \tilde{\phi}$ . However, when  $\phi = \tilde{\phi}$ , we have  $\langle f(x), g(x) \rangle = 1$ . Therefore sinusoidals are orthogonal.

Now, the similarity between a generic function  $f(x)$  and a sinusoidal  $s(x, \omega, \phi)$  can be readily expressed by the respective inner product, i.e.:

$$\langle f(x), s(x, \omega, \phi) \rangle = \int_{-\infty}^{\infty} f(x) s(x, \omega, \tilde{\phi}) dx = 0 \quad (48)$$

The higher the obtained value, the more similar the function  $f(x)$  is to the sinusoidal. Actually, the inner product  $\langle f(x), s(x, \omega, \phi) \rangle$  corresponds directly to the



Fourier transform of  $g(x)$ , which provides an effective manner of expressing  $f(x)$  as a linear combination of sinusoidals, which are more effectively handled as complex exponentials.

A particularly important property of the Fourier, as well as other orthogonal transforms is that, given the orthogonality between the basis functions, the calculation of the coefficients can be made independently one another.

Having obtained the transformation coefficients for every possible sinusoidal, it is then possible to recover the original function  $f(x)$ .

Now, we develop an analogous construction regarding the set operations incorporated into mfunctions.

First, we start by defining the mfunctional corresponding to the integral of the intersection between any two functions  $f(x)$  and  $g(x)$ . It is proposed here that this can be done as follows:

$$\ll f(x), g(x) \gg = \int_{-\infty}^{\infty} s_f s_g \min(s_f f(x), s_g, g(x)) dx \quad (49)$$

where  $s_f$  and  $s_g$  are the signs of  $f(x)$  and  $g(x)$ , and  $p(x) = f(x) \circ g(x) = s_f s_g \min(s_f f, s_g, g)$  is henceforth called the *mproduct* of  $f(x)$  and  $g(x)$ . In order to distinguish between the traditional and set-based inner products, we will refer to the latter as *common product*.

It is also possible to define other functionals, such as:

$$f(x) \circledast g(x) = \int_{-\infty}^{\infty} \max(s_f f(x), s_g, g(x)) dx \quad (50)$$

which is analogous to the outer product, being henceforth called the *sup product*.

This integral can be verified to effectively correspond to the common area between the two functions, with the area being taken considering the functions signs as in the above expression.

Thus, the Jaccard index for mfunctions can now be expressed as:

$$\mathcal{J}(f(x), g(x)) = \frac{\ll f(x), g(x) \gg}{f(x) \circledast g(x)} \quad (51)$$

As such, the Jaccard index can be understood as corresponding to the common product, which quantifies the similarity between the two msets or mfunctions in terms of the intersection-based common product (a functional), normalized relatively to the union of the mfunctions.

Let's illustrate this mfunctional respectively to sinusoidals. For simplicity's sake, we shall consider that all functions share the same support  $[0, T]$ ,  $T = 1f$ .

In case  $f(x) = g(x) = \sin(\omega x)$ , we immediately have that their intersection is  $f \cap g = g$ , from which

$$\begin{aligned} & \ll f(x), g(x) \gg = \\ & = \int_0^{T/2} \sin(\omega x) dt + \int_{T/2}^T \sin(\omega x) dt = 1 \end{aligned}$$

Now, in case  $f(x) = -g(x) = \sin(\omega x)$ , we have  $f(x) \cap g(x) = 0$ , implying

$$\begin{aligned} & \ll f(x), g(x) \gg = \\ & = - \int_0^{T/2} \sin(\omega x) dt - \int_{T/2}^T \sin(\omega x) dt = -1 \end{aligned}$$

For  $f(x) = \sin(\omega)$  and  $g(x) = \cos(\omega)$ , it follows that:

$$\begin{aligned} & \ll f(x), g(x) \gg = \\ & = \int_0^{T/4} \sin(\omega x) dt - \int_{T/4}^{T/2} \sin(\omega x) dt + \\ & + \int_{T/2}^{3T/4} \sin(\omega x) dt - \int_{3T/4}^T \sin(\omega x) dt = 0 \end{aligned} \quad (52)$$

So far, the common product between the sinusoidals has presented a remarkable similarity with the traditional inner product, being identical for the three situations above. In addition, we can conclude that the sine and cosine are orthogonal also regarding the common product.

However, let's now consider that  $f(x) = \sin(\omega)$  and  $g(x) = \sin(\omega + \phi)$ , with  $\phi \neq 0$ . Unlike the traditional inner product, the common product of these two functions will no longer be zero, being nevertheless comprised in the interval  $(0, 1)$ . As a consequence, except for the case  $f(x) = \sin(\omega)$  and  $g(x) = \cos(\omega)$ , no other combination of sinusoidals will result orthogonal one another.

Nevertheless, it is still possible to define transformations capable of decomposing (and recovering) a generic mfunction  $f(x)$  in terms of a combination of basic mfunctions such as sinusoidals. This can be readily accomplished as follows.

Let  $g_i(x)$ ,  $i = 1, 2, \dots, N$ , be the basis function of the transformation, and  $f(x)$  be the function to be analyzed. The transformation coefficients can be obtained as:

$$\begin{aligned} & \text{for } i = 1, 2, \dots, N \\ & \mathbf{c}_i = \ll f(x), g_i(x) \gg = \\ & = \int_{-\infty}^{\infty} s_f s_{g_i} \min(s_f f(x), s_{g_i}, g_i(x)) dx; \\ & f(x) = f(x) - g_i(x) \end{aligned}$$

Each of these coefficients corresponds to the effectively shared area between each pair of functions.

In the case of sinusoidal mfunctions, because they are generally not orthogonal, we will have that the above coefficients will depend on the order in which the basis functions are applied, so that several distinct transformations can be derived for a same function and a same basis. However, it is possible to employ optimization methods in order to achieve specific objectives.

The original function  $f(x)$  can then be recovered to a given precision (depending on the number and type of

basis functions) as:

$$\begin{aligned} f(x) &= 0; \\ \text{for } i &= 1, 2, \dots, N \\ f(x) &= f(x) + c_i g_i(x) \end{aligned} \quad (53)$$

## 10 The Cosine Similarity from the Multiset Perspective

Given two vectors  $\vec{X}$  and  $\vec{Y}$  with  $N$  elements each, their cosine similarity is commonly expressed as:

$$\text{Cs}(\vec{X}, \vec{Y}) = \text{Cos}(\theta) = \frac{1}{|\vec{X}||\vec{Y}|} \sum_{i=1}^N X_i Y_i \quad (54)$$

where  $\theta$  is the smallest angle between the two vectors. Typically the cosine similarity assumes all the elements to be nonnegative.

Now, we have already seen that vectors and multisets are closely related, in the sense that the latter can be derived from the former. Consequently, the cosine similarity index can be immediately understood in terms of multisets.

Let  $X$  and  $B$  be the multisets derived from  $\vec{X}$  and  $\vec{Y}$ , therefore sharing the same support with  $N$  elements. Then, we can write:

$$\text{Cs}(X, Y) = \frac{1}{\sum_{i=1}^N X_i \sum_{i=1}^N Y_i} \sum_{i=1}^N m(X_i) m(Y_i) \quad (55)$$

Therefore, the cosine similarity index can be understood as a normalized version of the Jaccard index where the union between the multisets has been replaced by their respective product, normalized by the product of the number of elements in the two multisets.

The above reasoning extends directly to two real mfunctions  $f(x)$  and  $g(x)$ :

$$\begin{aligned} \text{Cs}(f(x), g(x)) &= \\ &= \frac{1}{\int_{\Phi} |f(x)| dx \int_{\Phi} |f(x)| dx} \int_{\Phi} f(x) g(x) dx \end{aligned}$$

Now, it becomes possible to adapt the cosine similarity to intersection-based similarity by adopting the mproduct proposed in Section 9, i.e.:

$$\begin{aligned} \text{Cs}(f(x), g(x)) &= \\ &= \frac{1}{\int_{\Phi} |f(x)| dx \int_{\Phi} |f(x)| dx} \int_{\Phi} s_f s_g \min(s_f f(x), s_g g(x)) dx \end{aligned}$$

## 11 Toward Hybrid Signal Processing

The definition of the set common product as an analogue counterpart of the inner product binary operation be-

tween two mfunctions paved the way to the development of a field that we could denominate *set signal processing*, namely the handling and modification of signals by taking into account set operations including union and intersection *together* with all other algebraic operations such as sum, subtraction, product, division, etc. of mfunctions. The resulting set of concepts and methods can then constitute what is henceforth called *hybrid signal processing*. This interesting subject is developed in the current section.

We have already seen that the set common product plays a role analogous to the inner product. This enables us to define the *mconvolution*, namely the convolution between two multifunctions (also extended to other msets) as:

$$f(x) \diamond g(x)[y] = \int_{-\infty}^{\infty} \ll f(x) g(x-y) \gg dx \quad (56)$$

with  $y \in \mathbb{R}$ . Observe that, as with the traditional function convolution, the mconvolution is commutative.

It can be readily verified that the mconvolution quantifies the intersections areas between the functions, suitably adapted to take into account negative multiplicities.

Because the mconvolution does not penalize the difference between the two functions at a given position  $y$  with respect to the areas that are not overlapped, it becomes interesting to define a *set similarity mconvolution*, henceforth referred to as *sconv*, as follows:

$$f(x) \square g(x)[y] = \int_{-\infty}^{\infty} \frac{\ll f(x) g(x+y) \gg}{f(x) \circledast g(x-y)} dx \quad (57)$$

which, interestingly, is precisely the same as:

$$f(x) \square g(x)[y] = \int_{-\infty}^{\infty} \mathcal{J}(f(x), g(x)) dx \quad (58)$$

Observe that, since we are using  $x - y$  as argument of  $g()$  in the two above equations, the respective binary operations are more related to the correlation between two functions than to the convolution.

Thus, in case a more strict match is desired, the similarity convolution should be used instead of the mconvolution. Observe that the above approach extends immediately to scalar fields of any dimension as well as to generalizations of the Jaccard index as those described in [16].

The definition of these mconvolutions paves the way to theoretical and applied developments in all areas that already employ the traditional function convolution, including filtering, control theory, signal and image analysis, template matching, to mention but a few possibilities. This potential is illustrated with respect to template matching in the following.

The operation of *template matching* consists of, given a function  $f(x)$ , to quantify, along  $X$ , the similarity between its portions and another reference template function  $g(x)$ . This can be immediately implemented by by applying the mconvolution or similarity mconvolution to those two functions. High resulting values indicate portions of  $f(x)$  that are similar to  $g(x)$ .

Figure 7 presents the result of matching the template in (b) with the function in (a) by using mconvolution (c) and similarity mconvolution.

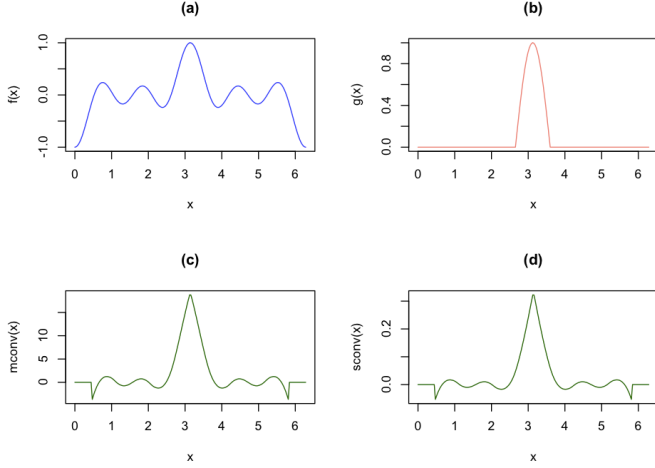


Figure 7: Template matching through mconvolution and sconvolution. The templated in (b) is to be compared with the mfunction in (a). This can ben effectively achieved by using the mconvolution between these two functions, whose result is shown in (c). The sconvolution (actually the Jaccard index for continuous functions) can also be employed in case a more strict quantification of the local similarity is being sought (d).

## 12 Generalized Multisets and Logic

Set operations are naturally associated to Boolean expressions. For instance, given three sets  $A$ ,  $B$ , and  $C$ , consider:

$$\begin{aligned} \text{Dataset Domain} &\iff \text{Model Domain} \\ A &= B^C \iff m_A = \neg m_B \\ A &= B \cup C \iff m_A = m_B \vee m_C \\ A &= B \cap C \iff m_A = m_B \wedge m_C \\ A &= B - C = A \cap B^C \iff m_A = m_B \wedge \neg m_C \end{aligned}$$

where  $m_A$ ,  $m_B$  and  $m_C$  are models associated to the datasets  $A$ ,  $B$ , and  $C$ .

It is this intrinsic association, so natural to humans to the point of sometimes not being realized, that often pro-

vides the basis for developing new models (logical combinations of models) while considering the adherence between respective datasets.

The generalization of the concept of multisets to several mathematical structures, as well as the identification of their relationship with several types of similarity indices, paves the way to applying multisets and mfunctions effectively in modeling activities.

As sets are naturally expanded to virtually every other mathematical structure, it becomes a topic of particular interest to identify which are the logical constructs corresponding to multiset operations such as sum, subtraction, product and division. Probably, these logic concepts have to do with the incorporation of the multiplicities implied by multisets into the logic reasoning, suggesting a logic with weighted or graded Boolean variables. For instance, when we say “two apples united with three pears” (multiset union), which has the direct logic meaning of “apples and pears”, when transformed to “two apples plus three pears” (multiset sum) would mean “two apples and three pears”.

The three levels of purported relationships between two sets, propositions, or functions can be summarized as in Table 1.

logical	$\neg$	$\wedge$	$\vee$
set theoretical	$\sqcap^C$	$\cap$	$\cup$
algebraic	$-$	$*$	$+$

Table 1: The purported interrelationships between three basic operations at the logic, set, and algebraic levels. .

## 13 Multisets in Pattern Recognition

The possibility to use msets to represent any type of density paves the way to interesting applications in pattern recognition and deep learning (e.g. [7, 8, 9, 10, 11]). In this section we illustrate how msets and the Jaccard index can be readily applied in order to quantify the similarity between two (or more) clusters represented by respective density functions.

Let’s consider the three sets of points in the scatterplot shown in Figure 8, which corresponds to the three species of iris flower in the frequently adopted iris dataset. Only two out of their 4 features have been chosen in the following example for simplicity’s sake.

The density obtained from the respective discrete samples through gaussian kernel expansion are shown in Figure 9.

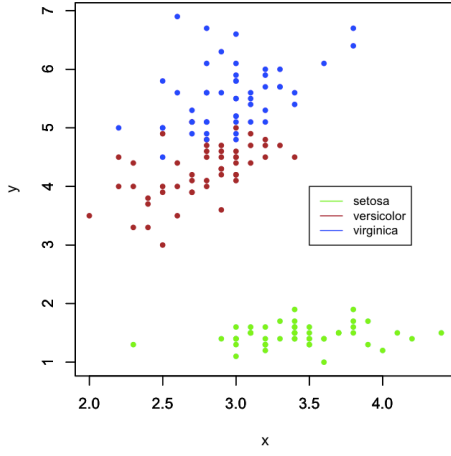


Figure 8: A scatterplot representing the distribution of three types of iris flowers represented by two respective features  $x$  and  $y$ .

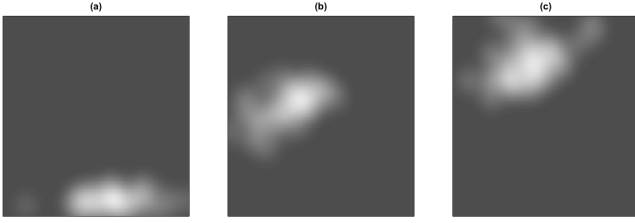


Figure 9: The three density scalar fields obtained by gaussian kernel expansion of each of the three categories.

The obtained multiset Jaccard index for each pairwise combination of categories are presented in Table 2.

	setosa	versicolor	virginica
setosa	1	2.6e-5	0
versicolor	2.6e-5	1	0.145
virginica	0	0.145	1

Table 2: The Jaccard indices obtained for pairwise combinations between the three iris species. The resolution has been limited to 6 digits.

The obtained results are fully compatible with the interrelationships between the three densities, or clusters, in Figure 8. In addition, the threewise Jaccard index from Equation 45 result nearly null, indicating a really small chance that the three densities correspond to the same cluster.

## 14 Concluding Remarks

The fascinating subject of multisets has been presented in a hopefully introductory manner.

We started with a brief review of traditional sets and their properties, which was followed by a progressive pre-

sentation of multisets and their characteristics. The possibility of obtaining multiset generalizations capable of dealing with functions, scalar fields, and densities, was then described and illustrated.

In addition to introducing several of the basic mset concepts, the present work also proposed how the universe mset can be defined in a robust manner by allowing the subtraction of msets to take negative values. This paved the way for recovering several properties analogous to traditional sets involving the complement operation, including the De Morgan theorem.

The extension of msets to mfunctions, namely traditional real functions represented as msets, was also proposed, paving the way to defining mfunctionals, of which the Jaccard index for mfunctions is one example, including operation analogous to the traditional inner product, but involving set operations, which paved the way to defining respective transformations analogous to the Fourier transform, though devoid of orthogonality.

Having defined an operation analogous to the inner product immediately enabled us to propose binary operators that are mset and mfunction counterparts of the traditional convolution between two functions, paving the way to achieving a field that can be called *hybrid signal processing* that is characterized by the incorporation of mset and mfunction counterparts to most of the concepts and operators adopted in signal processing.

The extension of the Jaccard index, which is intrinsically related to set theory, to msets and also to allow the consideration of more than two sets were also presented, which paves the way to employing these combined concepts for the characterization of relationships between clusters in feature spaces, a problem that is common to both the areas of pattern recognition and deep learning.

The presented concepts and methods can lead to several interesting applications, also motivating further integrations between the structures and properties between the domains of set theory, propositional logic, and analysis.

### Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

## References

- [1] J. Hein. *Discrete Mathematics*. Jones & Bartlett Pub., 2003.
- [2] D. E. Knuth. *The Art of Computing*. Addison Wes-

ley, 1998.

- [3] W. D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic.*, 30:36—66, 1989.
- [4] W. D. Blizard. The development of multiset theory. *Modern Logic*, 4:319?352, 1991.
- [5] P. M. Mahalakshmi and P. Thangavelu. Properties of multisets. *International Journal of Innovative Technology and Exploring Engineering*, 8:1–4, 2019.
- [6] D. Singh, M. Ibrahim, T. Yohana, and J. N. Singh. Complementation in multiset theory. *International Mathematical Forum*, 38:1877–1884, 2011.
- [7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2000.
- [8] K. Koutrombas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2008.
- [9] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.
- [10] J. Schmidhuber. Deep learning in neural networks:an overview. *Neural networks*, 61:85–117, 2015.
- [11] H. F. de Arruda, A. Benatti, C. H. Comin, and L. da F. Costa. Learning deep learning. Researchgate, 2019. [https://www.researchgate.net/publication/335798012\\_Learning\\_Deep\\_Learning\\_CDT-15](https://www.researchgate.net/publication/335798012_Learning_Deep_Learning_CDT-15). [Online; accessed 22-Dec-2019.].
- [12] L. da F. Costa. Modeling: The human approach to science. Researchgate, 2019. [https://www.researchgate.net/publication/333389500\\_Modeling\\_The\\_Human\\_Approach\\_to\\_Science\\_CDT-8](https://www.researchgate.net/publication/333389500_Modeling_The_Human_Approach_to_Science_CDT-8). [Online; accessed 1-Oct-2020.].
- [13] L. da F. Costa. An ample approach to modeling. Researchgate, 2019. [https://www.researchgate.net/publication/355056285\\_An\\_Ample\\_Approach\\_to\\_Data\\_and\\_Modeling](https://www.researchgate.net/publication/355056285_An_Ample_Approach_to_Data_and_Modeling). [Online; accessed 10-Oct-2021.].
- [14] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37:547–549, 1901.
- [15] Wikipedia. Jaccard index. [https://en.wikipedia.org/wiki/Jaccard\\_index](https://en.wikipedia.org/wiki/Jaccard_index). [Online; accessed 10-Oct-2021].
- [16] L. da F. Costa. Further generalizations of the jaccard index. [https://www.researchgate.net/publication/355381945\\_Further\\_Generalizations\\_of\\_the\\_Jaccard\\_Index](https://www.researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index), 2021. [Online; accessed 21-Aug-2021].